😺 PAWPULARITY 🐶

Team member:

Viacheslav Kiselev

Business understanding:

- ### Identifying your business goals:
  - Background: Millions of stray animals suffer on the streets or get euthanized by the shelters as they could not find them a new home. The lucky few strays who found a place in the shelter are in desperate need for a new home. One of the key factors in successful adoption is a cute and eye-catching photo, so a need for a software which will rate cuteness of the photo so a photographer can make/pick the best one.
  - Business goals: Create a tool which can be used both by shelters and general public which will help create better photos of pets looking for adoption
  - Business success criteria: If created system proves itself able to adequately rate the photos and be useful in the process of creating them

- ### Assessing your situation:
  - Inventory of resources: In order to build a successful model, the library of the images with their metadata used as training dataset should be continuously increased in size; Professional wildlife and household animal photographers may also be contacted in order to get information about important features of the photos and information on how to increase prediction accuracy in general; For faster and convenient neural network training high performance hardware is needed.
  - Requirements, assumptions, and constraints: Take a picture of an animal, feed it to the network; Assumptions: The animal in the photo is

visible and identifiable. Constraints: If an animal was not in the data set, the final score will be incorrect

  - Risks and contingencies: Risks: If the network encounters an animal which was not present in the training data set the score may be incorrect; Contingencies: If the animal is unknown, probably the score will be completely wrong and easily noticeable
  - Terminology: Pawpularity – predicted popularity/cuteness of a photo; Training data set – images used for neural network training
  - Costs and benefits: This tool will potentially save money for shelters as it will decrease the need for professional photographers taking the photos of the animals to get better outcome. Also, this project will potentially benefit anyone who wants to find a new home for a domestic animal

- Defining your data-mining goals:
  - Data-mining goals: We aim to create a highly accurate and confident score prediction model that will rate animal photos taken by people
  - Data-mining success criteria: If the created model will be able to confidently and precisely predict the Pawpularity scores. This will be tested on the Kaggle test dataset

Data understanding:

- Gathering data:
  - Outline data requirements: Images of household animals with metadata containing features and cuteness scores. At least 1000 annotated images are required
  - Verify data availability: Data for training was provided in decent amount on the Kaggle page. Training data set has 9 923 pictures in it

and all of them have annotations in a separate CSV file. Additional images with metadata are not easily available as Pawpularity score can be obtained only through the Petfinder app by uploading photos by yourself. Other features also need to be added manually

  - Define selection criteria: Images are of good enough quality for objects to be recognized. Image sizes are not too big so the training can be fast

- Describing data: The data will be taken from Kaggle PAWPULARITY competition. Training data set from Kaggle consists of 9 923 images of cats and dogs. All images have a unique ID and their own annotations, which can be divided into two classes features and score, stored in the CSV file. The features in the CSV file for each image are binary coded and the following features are present: Focus - Pet stands out against uncluttered background, not too close or far**,** Eyes – eyes are visible, Face – clear face, facing front or near front, Near – 50% of the photo is taken up by the pet, Action – actions is performed, Accessory – physical or digital accessory, Group – more than one animal, Collage – digitally edited photo, Human – human present, Occlusion - undesirable objects blocking part of the pet, Info - added text or labels, Blur - out of focus or noisy. Also, the data is unbalanced as some features appear more than the other.
- Exploring data: As the data set is unbalanced it will have to be balanced in order to ensure accurate prediction, but as getting more data to balance out the dataset is quite tricky it will have to be balanced out with removal of some data or by data augmentation during training
- Verifying data quality: Images are of good enough quality for objects to be recognized. Image sizes are not too big so the training can be fast. As mentioned before, the data set is unbalanced so in a situation where the balancing it out becomes impossible to balance it out the features can be removed leaving only the score and the photo.

Project plan:

- Balancing the train dataset, splitting it into train, test and evaluation datasets for neural – 5 Hr.
- Apply data augmentation (cropping, resizing, gamma correction and other parameters) on the training images – 10 Hr.
- Train the neural network on augmented and not-augmented datasets – 48 Hr.
- Evaluate the network on the evaluation dataset – 5 Hr.
- Tune hyperparameters, train again if needed and re-evaluate – 5 – 10 Hr.
- Test the model on the test data set from Kaggle – 1 Hr.

LINK: https://github.com/kiselevsl/DS2022_PAWPULARITY